DATA MANAGEMENT PLAN (2 PG MAX)

The sustainability of gene function prediction resources is very much contingent on the hardware and servers on which they are stored and run as well as the publicly available data used in model development and validation.

The existing infrastructure at Yale and Royal Holloway University of London has served investigators well for over a decade, but we aim to improve the current setup by making it reliable and robust for supporting all the proposed tools as well as more accessible to the scientific community. To this end will make use of new technologies such as cloud computing.

Specifically, we intend to use Amazon Web Services (AWS) for distributing the network analysis and function prediction tools, and intend to make use of the Amazon Elastic Compute Cloud EC2 (processing) and S3 (storage). AWS EC2 enables flexible, resizable online resources, and would serve as a sensible means for distributing the developed services, as it provides high performance computing (thereby delivering faster analyses of complex biological networks), processing resources which adjust to user demand, reliability, and greater security, with access privileges we define (e.g., we may enable the users who submit their customized biological network, gene expression and phenotype data; alternatively, we will provide a number of carefully curated datasets and access and links to other publically available data).

Here, we summarize various components of the proposed resources, along with the means by which we intend to disseminate each:

- **Source code**, as used in constructing the various software components will be made available through open access repositories, such as sourceforge, github, or google code.
- **Web-services** (network analysis and phenotype function prediction tools): each of the servers would be encapsulated and made available as a virtual machines (see below for a description and the advantages of virtual machines), which may be downloaded from our servers, and then stored locally by the user.
- **Databases**: a variety of curated biological networks used in algorithm testing and validation will be bundled up into a single virtual machine, and distributed through our host websites and open access repositories. We intend to move this dataset to AWS S3 for storage, and easy access.

As mentioned, the various servers, as well as the large dynamic datasets, would each be converted into a separate virtual machine (VM). A VM encapsulates an entire piece of software (even as large and complex as an entire operating system), and may easily be packaged up for easy storage and distribution. A VM behaves like an autonomous computer within a real computer. One primary advantage of using VMs is that they may be run on many different systems (such as Windows, Linux, or OSX). Upon execution, a VM has the appearance of booting a new computer and operating system (a process called virtualization), and it is this process which precludes the inconvenience and needed time of having to install and configure software on the host's operating system, thereby enabling developers (in this case, us) to build autonomous, customized, turnkey, ready-to-go execution environments with a complete operating system and all required software packages, libraries, and data files packaged into a single VM. These properties make VMs suitable as a vehicle for distributing our software tools through cloud-based systems, and this makes their use much easier for investigators. Users would run the VMs locally, and for large-scale deployment, they may be exported to AWS EC2. This encapsulation strategy not only provides our developed tools as a set of discrete products for the end user, but also contributes to the sustainability of the resource itself.

Cristina Sisu 1/8/14 16:14
Deleted: Loregic,

Cristina Sisu 1/8/14 16:14
Deleted: P

Cristina Sisu 1/8/14 16:14
Deleted: F

Cristina Sisu 1/8/14 16:14
Deleted: P

Cristina Sisu 1/8/14 16:14
Deleted: T

Cristina Sisu 1/8/14 16:15
Deleted: the

Cristina Sisu 1/8/14 16:15
Deleted: regulatory

Cristina Sisu 1/8/14 16:15
Deleted: dataset

The optimum development of the proposed resources is also depended on the availability of scientific data used to test and validate the methods. As part of numerous consortia, Dr Gerstein has unrestricted access and to a large variety of functional genomics and network data for human and numerous other model organisms. In this project we are going to make use of gene expression and regulation data from:

Cristina Sisu 1/8/14 16:22
Deleted: function prediction

- ENCODE (www.genome.gov/encode/)
- modENCODE (www.modencode.org)
- the Cancer Genome Atlas (http://cancergenome.nih.gov)

available for *H. sapiens, C. elengans, S. cerevisiae, D. melanogaster, S. Pombe, M. Musculus, D. rerio.*

The phenotype prediction tools will leverage on data available from phenotype ontologies that are publicly accessible through the online resource PhenomicDB (www.phenomicdb.de).